EMERGING MARKETS
JOURNAL
ISSN 2159-242X (print) ISSN 2158-8708 (online)

HOW TO GROUP FINANCIAL DATA WITH MAXIMUM HOMOGENEITY?

Mehmet Baran
Marmara University
e-mail: mehmet.baran@marmara.edu.tr

Sıtkı Sönmezer – corresponding author
Beykent University
e-mail: sitkisonmezer@beykent.edu.tr

Volume 3 (2013) | ISSN 2158-8708 (online) | DOI 10.5195/emaj.2013.36 | http://emaj.pitt.edu

Abstract

Grouping may be an obstacle itself or it may have to be improved to extract better information out of a data stream. Finding trends and dividing a population into parts may be crucial for analyses. This paper offers a modified version of Fisher method that may smoothen the cut point transitions and give out better results. Proven methodology is given with a comparison with the original method. The method may be helpful in forming subgroups in financial data, possibly in technical analyses.

Keywords: Grouping, Fisher Method, Trends, Cut points

ULS D-Scribe digital publishing

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program, and is cosponsored by the University of Pittsburgh Press

Volume 3 (2012) | ISSN 2158-8708 (online) | DOI 10.5195/emaj.2012.36 | http://emaj.pitt.edu

How To Group Financial Data With Maximum Homogeneity?

Mehmet Baran, Sıtkı Sönmezer

INTRODUCTION

It is often the case, when trying to extract information out of a financial time series data; researchers face the problem of dividing the financial data into homogenous parts. Alternatively, when the researchers aim to determine trends within a data set, it may be essential to know how to decompose a sample into sub- samples in which homogeneity is maximized. Intuition or segmenting the data visually may be used for grouping the data, however; these methods lack the efficiency when compared with the statistical methods that are numerous in the literature.

Various disciplines have the problem of clustering a set of objects as there are different usages, applications and objective functions, thus, there are many variations of this clustering problem. Maharaj and Inder (1999) and Duncan, Gorr and Szczypula (2001) have given methods of clustering time series data. Therefore, numerous clustering algorithms from different fields of study are present in the literature however; the merits of these algorithms are dubious (Gonzales 1985).

The method developed in this study is a modification of the method suggested by Fisher (1958) that tries to find a methodology for grouping and forming sub-samples. It has to be noted that Fisher considered two related, but different problems; first one is introduced as the unrestricted problem and second as the restricted problem. This paper deals with the restricted problem part of the method, which arises in many cases in finance where certain conditions, on the basis of prior information, theory or for convenience, are imposed most of the time and offers a modified version that gives smaller error terms. The first part of the study provides literature survey regarding with the method used in finance related sectors. Secondly, Fisher method and how it is improved is presented in the second part and the final part is concluded with an example showing the improvement.

LITERATURE REVIEW

Despite defining homogenous groups are studied vastly by statisticians and hydrologists, in this paper, only the articles that have utilized methods that determine homogeneous grouping with financial data are examined. However, one main method, Fisher method, has to be mentioned as it has been used by many studies, including this one. The method introduced by Fisher (1958) attempts to minimize the MSE, takes the decrease n variance into consideration due to combining which decreases variance and ignores the increase in bias. A review of the algorithm is presented here below.

Huang and Zhang have investigated structural changes in Singapore’s private housing market and in particular the impact of government policies on housing price determination by utilizing Fisher method to find “1) the specification of the number of changes in the model 2) the detection of the change point, or the boundaries of intervals over which each of the model pieces applies 3) the estimation of the model parameters within each subdomain.” (2012)

Kumar and Patel (2010) have come up with a new method derived from Fisher method which is based on the tradeoff between decreased variance and increased bias arising from combining. With this modified version, it is possible to determine when it is beneficial to combine or not. Adams and Lim also utilized the Fisher method in forming sub groups for their data to minimize sum of squared deviations for growth polarity (2011).

REVIEW OF THE CLASSICAL FISHER GROUPING ALGORITHM

In order to understand how Fisher grouping algorithm works, consider that a time series consisting of *N* data points *x(i), 1<=i<=N*, is given. Assume that this time series is divided into K contiguous and mutually exclusive segments:

$$[1, N_1] \cup [N_1 + 1, N_2] \cup \dots \cup [N_{K-2} + 1, N_{K-1}] \cup [N_{K-1} + 1, N] \tag{1}$$

where $[a, b]$denotes all the integers between *a* and *b*, inclusive, and $N_r > N_s$if *r>s*. Let the mean for the *j* th segment $[N_{j-1} + 1, N_j]$be denoted by M_j:

$$M_J = \frac{1}{N_j - N_{j-1}} \sum_{i=N_{j-1}+1}^{N_j} x\,(i) \tag{2}$$

Note that $N_0 = 1$and $N_k = N$. The mean squares (MS) error term for the *j*th segment, e_j can be computed as:

$$e_j = \sum_{i=N_{j-1}}^{N_j} \left(x(i) - M_j\right)^2 \tag{3}$$

and the MS error for the whole time series, *E*, is defined as the sum of the MS errors of all segments

$$E = \sum_{i=0}^{K-1} e_i \tag{4}$$

After these preliminary definitions, Fisher grouping problem can be defined as follows: Determine the segment boundaries N_1,N_2, ... ,N_{K-1}in (1) in such a way that the MS error *E* defined in Eq. (4) is minimized. The set of error minimizing segment boundaries $(N_1, \dots, N_{k-1})$is called the Fisher segmentation. Fisher segmentation gives the most homogeneous possible division of any given time series into *K* segments. The segment boundaries represent the transition regions between neighboring segments.

It is possible to solve the Fisher grouping problem via exhaustive search only for small time series. The number of all possible K-segmentations for a time series of length N is

$$C(N,K) = \frac{N!}{K!\,(N-K)!} \tag{5}$$

Which grows exponentially with *N*. Hence, exhaustive search is impractical for large *N*. Fortunately, a dynamic programming approach is provided by Fisher. This approach consists of three main steps which is listed herebelow:

1) Calculate the error matrix *e[i,j]*

$$m[i,j] = \frac{1}{j-i} \sum_{q=i+1}^{j} x\,(q) \tag{6}$$

$$e[i,j] = \sum_{q=i+1}^{j} (x(q) - m[i,j])^2 \tag{7}$$

Where *m[i,j]* denotes the mean of all points within *[i+1,j]*, and *e[i,j]* denotes the variance of all points in the same interval.

2) Compute the vector *D[2,r]* for all *r*, *r<N* as the cost of the optimum two-segment division of the interval $[1, r]$

$$D[2,r] = min_k(e[1,k] + e[k+1,r]) \tag{8}$$

3) Lastly, for *m>2*, compute

$$D[m,r] = min_k(D[m-1,k] + e[k+1,r]) \tag{9}$$

Note that *D[m,r]* denotes the optimal LS cost of the division of time series *[1,r]* into *m* segments.

Given the above algorithm, one can compute D[K,N] recursively, which is the quantity we are looking for. At the first step, one must calculate the matrix *a[i,j]*. Then the vectors *D[2,2]*, *D[2,3]*, ..., *D[2,N]* must be computed. Then using these values, one can compute *D[3,3]*, *D[3,4]*, ..., *D[3,N]* and so on. The algorithm will stop when it reaches *D[K,N]*, which is the desired quantity.

All steps of the algorithm has $O(N^2)$ complexity, and there are *K* steps. Hence, the overall complexity is $O(KN^2)$.

PROBLEMS OF FISHER GROUPING IN SEGMENTING FINANCIAL DATA

Fisher grouping is perfect for time series where the data remains in near-constant regimes for long durations, and the transitions between these near constant regimes are relatively fast. This is one reason why Fisher algorithm found particular favor with hydrologists: Most of the lakes and rivers have near-constant depth for long periods of time, with very rapid change of depth during short intervals of time (ie, spring rains) between them. Unfortunately, financial data does not fit into this pattern. "Trends" in financial data are rarely defined by approximately constant levels. More frequently, trends are defined by approximately constant rates of growth or decay, and changes in the growth rate can be considered as changes in trend. Therefore, Fisher algorithm in its classical form (ie, Equations *(1)-(9)* given above) is largely unsuitable for discovering trends in financial data, and boundaries between neighboring trends.

Below, a more developed version of Fisher's algorithm will be proposed, in which the deficiencies mentioned will be addressed. Fisher's original grouping algorithm is "zero order", as it approximates segments with constants, which are zero order curves. The proposed algorithm, in contrast, will approximate segments with first order curves, i.e. lines, which are ordinary least squares (OLS) approximations of the data in the segment. We will call this newly proposed algorithm as the first order Fisher grouping algorithm.

FIRST ORDER FISHER GROUPING ALGORITHM

The main difference between zeroth order and the proposed first order algorithms comes from a modification of Eqs. (6)-(7), which calculates the segment error. In Fisher's original article, this error is calculated around the segment mean , via Eqns. (6)-(7). In contrast, we propose the following equations for segment error calculation:

$$n^N[i,j] = \sum_{k=i+1}^{j} kx\,(k) - \frac{1}{j-i} \sum_{k=i+1}^{n} k \sum_{k=i+1}^{n} x\,(k) \tag{10}$$

$$n^D[i,j] = \sum_{k=i+1}^{j} k^2 - \frac{1}{j-i} \left(\sum_{k=i+1}^{n} k \right)^2 \tag{11}$$

$$n[i,j] = \frac{n^N[i,j]}{n^D[i,j]} \tag{12}$$

$$m[i,j] = \frac{1}{j-i} \sum_{k=i+1}^{j} x\,(k) - \frac{n[i,j]}{j-i} \sum_{k=i+1}^{j} k \tag{13}$$

$$e[i,j] = \sum_{k=i+1}^{j} (x[k] - m[i,j]k - n[i,j])^2 \tag{14}$$

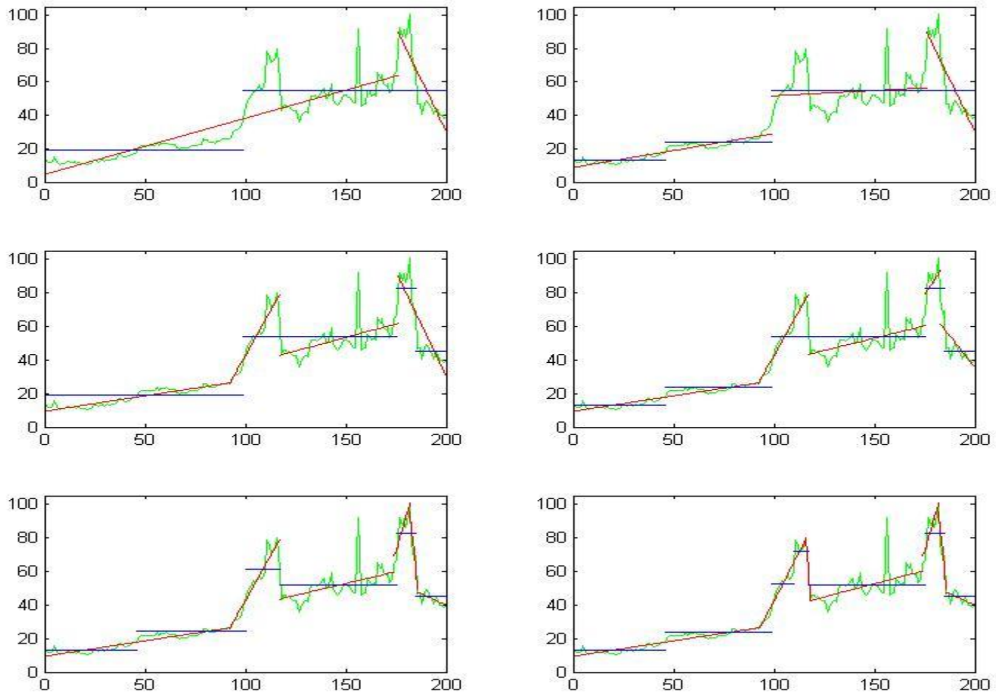
Note that these are the well-known linear regression equations. After calculating the *e[i,j] matrix*, the rest of the first order algorithm is exactly the same with the zeroth order algorithm. Namely:

1) Calculate *e[i,j]* by using Eqs. (10-14)
2) Calculate *D[2,r]* for all *2<r<K* by using Eq. (8)
3) Calculate *D[m,r]* by using Eq. (9) till *D[K,N]* is found.

As will be seen in the next section, even though only the first step of the segmentation algorithm is changed, the resulting segmentations are dramatically different

Application To Financial Data

In the figures below, the green curve represents the time series for the closing price of fener TI equity for 3000 days, subsampled to 200 data points. Blue curves represent the trends discovered by the classic zero-order Fisher algorithm. The red curves represent the trends discovered by the first order Fisher algorithm, which is proposed in this article. From the upper-left figure to the lower-right figure, progressively more trends are searched in the data. In the upper-left figure, data is divided into two trends. This is increased to seven trends in the lower right figure.



It is clear from the figures that the two algorithms give markedly different segmentations on the same data set. As predicted above, classic Fisher algorithm is particularly weak when the time series diplay a uniform rate of increase for long periods of time. This is obvious at the lower right figure where data is divided to seven trends. At the range [0, 110] there is clearly two trends in data, and first order algorithm clearly discovers them. Zero order algorithm discovers five trends at the same range. This is the best the zero- order algorithm can do, as it can only approximate uniformly increasing data by staircase-like constant segments. Same phenomenon can be seen, in a less extreme form, in all the other plots.

Conclusion

Our results show that classic zero order Fisher algorithm is not the ultimate solution for partitioning financial data into trends and there is room for improvement. First order Fisher algorithm, which is proposed in this article, succeeds in increasing the accuracy of the classical algorithm considerably, while not adding to its computational complexity. It gives practically the same segmentations with the zero order algorithm when the data remains constant over long durations, but gives markedly better segmentations when the data increases at a constant rate.

It is possible to extend the first order algorithm to second order to handle quadratic increases, or to impose a continuity condition between neighbouring segments. We plan to extend our research into these directions in forthcoming publications.

References

Duncan, G., Gorr, W. L., & Szczypula, J., 2001, Forecasting analogous time series. In J. S. Armstrong (Ed.), Principles of forecasting: A handbook for researchers and practitioners, 195–213. Dordrecht: Kluwer.

Fisher W. D., 1958, On grouping for maximum homogeneity. Journal of the American Statistical Association, 53, 789–798.

Huang W. & Y. Zhang, 2012, Structural Change Modeling of Singapore Private Housing Price in Simultaneous Equation Model, Journal of Financial Risk Management. Vol.1, No.2, 7-14

Kane J. A.and J. J. Lim, 2011, Global Growth Poles in a Multipolar World Economy

Kumar M, Nitin R. Patel, 2010, Using Clustering to Improve Sales Forecasts In Retail Merchandising, Annals of Operations Research, February, Volume 174, Issue 1, 33-46

Maharaj, E. A., & Inder, B. A.,1999, Forecasting time series from clusters ,Monash Econometrics and Business Statistics Working Papers, Monash University, Department of Econometrics and Business Statistics, 9/99.

Teofilo F. Gonzales, 1985, Clustering to Minimize the Maximum Inter-cluster Distance, Theoretical Computer Science 38, 293-306